



AEC Computing and Applied Mathematics Center

AEC RESEARCH AND DEVELOPMENT REPORT

TID-4500
13th Ed.

NYO-8674
PHYSICS

A NON-RANDOM SAMPLING METHOD, BASED ON
CONGRUENCES, FOR "MONTE CARLO" PROBLEMS

by

R. D. Richtmyer

September 26, 1958

Institute of Mathematical Sciences

NEW YORK UNIVERSITY

NEW YORK, NEW YORK

NYO-8674

NYO-8674



This report was prepared as an account of Government sponsored work. Neither the United States, nor the Commission, nor any person acting on behalf of the Commission:

- A. Makes any warranty or representation, express or implied, with respect to the accuracy, completeness, or usefulness of the information contained in this report, or that the use of any information, apparatus, method, or process disclosed in this report may not infringe privately owned rights; or
- B. Assumes any liabilities with respect to the use of, or for damages resulting from the use of any information, apparatus, method, or process disclosed in this report.

As used in the above, "person acting on behalf of the Commission" includes any employee or contractor of the Commission to the extent that such employee or contractor prepares, handles or distributes, or provides access to, any information pursuant to his employment or contract with the Commission.

UNCLASSIFIED

AEC Computing and Applied Mathematics Center
Institute of Mathematical Sciences
New York University

TID-4500
13th Ed.

NYO-8674
PHYSICS

A NON-RANDOM SAMPLING METHOD, BASED ON
CONGRUENCES, FOR "MONTE CARLO" PROBLEMS
by

R. D. Richtmyer

September 26, 1958

Contract No. AT(30-1)-1480

INSTITUTE OF MATHEMATICAL SCIENCES
25 Waverly Place, New York 3, N. Y.

UNCLASSIFIED

ABSTRACT

The theory of systematic sampling, a possible competitor to random sampling for problems of Monte-Carlo type, is discussed. It is based on the asymptotic distribution, in the unit cube in a space of many dimensions, of a sequence of points generated by congruences (mod 1) of the multiples of a vector with independent irrational components. Theorems based on algebraic number theory, and ones based on a statistical treatment, concerning the rate of decrease of the error with increasing sample size, are given. The method is illustrated by application to a simple, multiplicative, stochastic problem. The error is seen to be substantially smaller than when the sequence of points in the unit cube is a random one. This is strongly suggested by the algebraic theorems and confirmed by the statistical theorems; however, it is likely that, in practical Monte Carlo problems of the usual complexity, the improvement over random sampling will be evident only for very large sample sizes.

NEW YORK UNIVERSITY
INSTITUTE OF MATHEMATICAL SCIENCES
LIBRARY
25 Waverly Place, New York 3, N.Y.

TABLE OF CONTENTS

	Page
Abstract.....	2
Section	
1. Introduction and Summary.....	4
2. The Monte Carlo Method.....	8
3. Theorems on Systematic Sampling.....	12
4. Statistical Treatment.....	20
5. Comments.....	24
6. Numerical Examples.....	32
Bibliography.....	41

NEW YORK UNIVERSITY
INSTITUTE OF MATHEMATICAL SCIENCES
LIBRARY
25 W. 4th St., New York 3, N. Y.

A NON-RANDOM SAMPLING METHOD, BASED ON
CONGRUENCES, FOR "MONTE CARLO" PROBLEMS1. Introduction and Summary.

During the past ten years, several studies have been made of systematic (non-random) sampling methods, for calculations of the Monte Carlo type, based on theorems of H. Weyl^[1] on the asymptotic distribution, in n-space, of a set of points generated by congruences. The object is to find a computing method, similar to Monte Carlo, in which the error of the approximation decreases more rapidly, with increasing sample size, than when random sampling is used. Although the method has not yet achieved success in practical applications, there seems to be enough interest in it to warrant giving an account.

In Section 2 it is shown that when the Monte Carlo method is applied to the stochastic problems of chain reactions, cosmic ray showers, and the like, the problem to be solved may be interpreted as that of evaluating a definite integral

$$(1) \quad I = \int_0^1 \cdots \int_0^1 f(x_1, \dots, x_k) dx_1 \dots dx_k$$

over the unit cube in a large number of dimensions. In sampling methods, the integral is approximated by choosing

a sequence of points $(x_1^{(n)}, \dots, x_k^{(n)})$ ($n=1, 2, \dots, N$), in the k -dimensional cube, and writing

$$(2) \quad I_N = \frac{1}{N} \sum_{n=1}^N f(x_1^{(n)}, \dots, x_k^{(n)})$$

as an approximation to I .

In the standard Monte Carlo method, these N points are chosen at random, in the cube; the methods of statistics can then be used to show that $I_N \rightarrow I$, as $N \rightarrow \infty$, in the sense that

$$(3) \quad \text{Expectation } (I_N) = \langle I_N \rangle = I ,$$

$$(4) \quad \text{Std. Dev. } (I_N) = \sqrt{\langle (I_N - I)^2 \rangle} \propto \frac{1}{\sqrt{N}} ,$$

which, for comparison with the non-random sampling method, we paraphrase by writing

$$(5) \quad I_N - I \approx \mathcal{O}\left(\frac{1}{\sqrt{N}}\right) ;$$

furthermore, standard sampling theory can be used to estimate the error of the sample mean I_N in terms of the sample variance (which is usually calculated concurrently with I_N , by the computing machine).

In the systematic sampling method, the sequence of points in the k -dimensional unit cube is chosen as follows:

Let $\xi_1, \xi_2, \dots, \xi_k$ be k linearly independent^{*}, real, irrational numbers. When these have been chosen, there are no further arbitrary elements in the method - the sequence of points is then fully determined: the coordinates of the n^{th} point are

$$(6) \quad (x_1^{(n)}, \dots, x_k^{(n)}) = (\{n\xi_1\}, \dots, \{n\xi_k\}) ,$$

where $\{x\}$ denotes the fractional part of x .

H. Weyl showed, first, that the sequence (6) is asymptotically uniformly distributed in the unit cube and, second, that when the points in (2) are chosen in this way,

$$(7) \quad I_N \rightarrow I \text{ as } N \rightarrow \infty ,$$

provided that the function $f(x_1, \dots, x_k)$ is Riemann-integrable. ^{**}

There is considerable evidence that the fluctuations of the sequence (6) may be substantially less than that of a random sequence. Section 3 summarizes this evidence, which consists of theorems by A. Ostrowski, by the writer, and by L. G. Peck. These theorems indicate that for a sufficiently smooth integrand $f(x_1, \dots, x_k)$, the approach of I_N to I is asymptotically more rapid than indicated by (5).

* There is no vanishing linear combination of them that has all rational coefficients.

**Weyl's first result means that $I_N \rightarrow I$ if $f(x_1, \dots, x_k)$ is the characteristic function of any rectangular subregion of the cube.

Unfortunately, the theorems referred to are not broad enough to be applicable to practical Monte Carlo problems because of the restriction to $k = 1$ in Ostrowski's theorem and to very smooth integrands in the theorems of the writer and of Peck. A new approach to the problem is therefore taken, in Section 4, by reintroducing statistical considerations. If, at the beginning of a problem, the numbers $\xi_1, \dots, \xi_k$ are chosen at random from some continuous distribution, then, with exception of choices having probability zero, they will indeed turn out to be linearly independent irrational numbers. We can therefore regard the $\xi_1, \dots, \xi_k$ as random variables (they will be taken to have a uniform distribution in the unit cube, since there is no loss of generality in this choice); for given N (and a given function $f(x_1, \dots, x_k)$), I_N is then a function of the random variables $\xi_1, \dots, \xi_k$, and we may discuss its statistics.

Our main result is that, with the points in (2) chosen according to (6), but with the $\xi_1, \dots, \xi_k$ now regarded as random variables, we have

$$(8) \quad \text{Expectation } I_N = \langle I_N \rangle = I ,$$

$$(9) \quad \text{Expectation } |I_N - I| = \langle |I_N - I| \rangle = \mathcal{O}\left(\frac{\ln N}{N}\right)$$

(see equation (31)), provided only that $f(x_1, \dots, x_k)$ is sufficiently smooth that its Fourier series is absolutely

convergent. This is a sufficiently great improvement over straight random sampling that, if the coefficients implied in (5) and (9) do not differ too vastly, the systematic sampling method should be quite worthwhile in practice. The requirement of absolute convergence of the Fourier series for $f(x_1, \dots, x_k)$ can often be achieved by means of sampling "tricks" such as the ones that have been used, in random sampling, to reduce the sample variance (this is illustrated in the numerical examples).

Section 5 contains various comments on the method; and Section 6 discusses a few numerical examples.

The improvement over random sampling appears to be rather small but may be of some slight practical interest.

2. The Monte Carlo Method.

In the usual Monte Carlo method of calculating the properties of a neutron chain reaction, one is concerned with a large collection or ensemble of neutron chains, each starting with a single neutron, and all developing under identical conditions, determined by the distribution of fissionable and scattering material. The chains differ from each other, even though they are all started in an identical manner, because the occurrence and characteristics of the nuclear events are determined by chance. The probability laws governing these events are presumed known and one wishes to determine the statistical properties of the ensemble of chains.

For this purpose, one constructs numerical descriptions, mathematically, of a large number of representative chains, by assigning arbitrary values to the various quantities that determine the sequence of events constituting the chains. The values are assigned at random, with probabilities distributed in accordance with the aforementioned probability laws. Specifically, the assignment is made in the following way: there is supposed given an inexhaustible supply of "random" numbers, i. e., an infinite sequence of numbers,

$$(10) \quad x_1, x_2, x_3, \dots, x_i, \dots$$

uniformly distributed in the interval $(0, 1)$, of which each is statistically independent of the numbers preceding it in the sequence. These numbers are used one by one and then discarded. When, in the course of constructing the numerical description of one of the chains, it becomes necessary to assign a value to a quantity y for which a probability distribution $P(y)$ is given (by this is meant: $P(y_0)$ is the probability of $y < y_0$ and, in particular, $P(-\infty) = 0$, $P(+\infty) = 1$), then the value assigned to y is y_0 given by the equation: $P(y_0) = x_i$, where x_i is the next unused number of the sequence (10). For computing purposes, this relation is expressed in the form:

$$(11) \quad y_0 = F(x_i),$$

where $F(x)$ is the inverse of the function $P(y)$. The construction of each chain is continued until all the events occurring up to a given time ("census time") have been determined. Then, the next chain is constructed, and so on.

(In practice, the calculations for different chains are often interspersed, the events being taken in more or less chronological order, regardless of which chain they belong to; but the final result is the same as that described above, except for an irrelevant permutation of the random numbers.)

When the descriptions of sufficiently many representative chains have been prepared in this way, average properties of the chain reaction can be obtained by statistical analysis of them. Suppose, for example, that the problem is to find the average value of the total kinetic energy E of all neutrons in the system at the end of a specified time after the introduction of the initiating neutron. We compute the value of E for each representative chain from its numerical description and average the result for all the chains.

The value of E (or of any quantity whose average value is desired) for a given representative chain depends on the values of those "random" numbers of the sequence (10) that entered into its description. That is, we can write,

$$(12) \quad E = f(x_i, x_{i+1}, x_{i+2}, \dots) ,$$

for some i . The function f must, in general, be regarded as a function of infinitely many variables, because one cannot tell in advance how many random numbers will be required for the description of a given chain up to the end of the specified time interval, unless one is willing to adopt some ruse such as assigning a minimum time between collisions for a neutron; for, otherwise, an indefinitely large number of events may, in principle, take place in the interval. If it turns out, for a given chain, when the values of the appropriate x 's are examined, that m random numbers suffice for determining the chain, then, for those particular values of the first m arguments in (12), the function f is independent of the values of the remaining, infinitely many, arguments.

In general, the function f is very complicated, and cannot be easily written down, but in any case it is clear that the objective of the Monte Carlo calculation is an approximate evaluation, by use of the random numbers, of the integral

$$(13) \quad \bar{E} = \int_0^1 dx_1 \int_0^1 dx_2 \dots f(x_1, x_2, \dots)$$

over the unit cube in a space of infinitely many dimensions. Each chain is associated with a point in this unit cube. $\bar{E}$ is the limit towards which the average of E must tend as the

number of mathematical representants in the ensemble tends to infinity, because the variables in (13) are random numbers uniformly and independently distributed in the interval $(0,1)$. The representative chains of Monte Carlo are associated with points randomly distributed in the unit cube.

In practice one can restrict consideration to a finite number k of dimensions, and this leads to an integral of type (1), mentioned in the first section. If k is sufficiently large (say 100 - 1000 in typical cases), the probability that the computing machine will ask for more than k "random" numbers, in the construction of a single chain, is so remote that this possibility can be ignored - i. e. one can postpone making any decision on what to do about it until it occurs.

3. Theorems on Systematic Sampling.

Ostrowski's results^[2] concern the case $k = 1$, so that

$$(14) \quad I = \int_0^1 f(x) dx, \quad I_N = \frac{1}{N} \sum_{n=1}^N f(\{\xi\}).$$

His results are stated for the function

$$(15) \quad f(x) = \left\{ \begin{array}{ll} 1, & \rho_1 < x < \rho_2 \\ 0, & \text{otherwise} \end{array} \right\},$$

where $0 \leq \rho_1 < \rho_2 \leq 1$; but, clearly, they also hold for

any function $f(x)$ of bounded variation. His main conclusions are twofold:

1. If ξ is a quadratic irrational, or any irrational whose continued fraction expansion

$$\xi = a_0 + \frac{1}{a_1} + \frac{1}{a_2} + \dots \quad (a_0, a_1, \dots \text{ are positive integers})$$

has bounded partial denominators, i. e., satisfies

$$a_j \leq A$$

for some fixed A , then

$$(16) \quad |I_N - I| \leq \text{constant} \frac{\ln N}{N}.$$

If $f(x)$ is as defined in (15), the constant is $\leq 36A$.

2. If ξ is an irrational number whose continued fraction expansion has partial denominators increasing sufficiently slowly so that $q_{i+1} = O(q_i^r)$, for fixed $r > 1$, where the successive best rational approximations to ξ are p_i/q_i , that is, where

$$q_0 = 1, \quad q_1 = a_1, \quad q_{i+1} = a_{i+1}q_i + q_{i-1},$$

then,

$$(17) \quad I_N - I = O\left(\frac{1}{N^{1/r}}\right)$$

This second conclusion of Ostrowski can be combined with the conclusion of Roth's theorem^[5] establishing the truth of the Thue-Siegel conjecture that, if ξ is any algebraic number and $\epsilon > 0$, then

$$\left| \xi - \frac{p}{q} \right| > \frac{1}{q^{2+\epsilon}}$$

for all integer pairs (p, q) , with at most a finite number of exceptions. Since, always,

$$\left| \xi - \frac{p_i}{q_i} \right| < \frac{1}{q_i q_{i+1}},$$

it follows that $q_{i+1} = O(q_i^{1+\epsilon})$; hence, if ξ is any algebraic number and $\epsilon > 0$, then

$$(18) \quad |I_N - I| = O\left(\frac{1}{N^{1-\epsilon}}\right),$$

where I_N and I are given by (14).

The theorem stated by the writer in [3] is the following: if $\xi_1, \dots, \xi_k$ are linearly independent irrationals in a real algebraic number field of degree n (where, necessarily, $n \geq k+1$), and if the function $f(x_1, \dots, x_k)$ is such that the coefficients of its Fourier series

$$f(x_1, \dots, x_k) = \sum_{-\infty}^{\infty} (q_1, \dots, q_k) a(q_1, \dots, q_k) e^{2\pi i(q_1 x_1 + \dots + q_k x_k)}$$

satisfy the inequality

$$(19) \sum_{(q_1, \dots, q_k)} \left\{ \max_{(j)} |q_j| \right\}^{n-1} |a(q_1, \dots, q_k)| < \infty ,$$

then,

$$I_N - I = O\left(\frac{1}{N}\right) \text{ as } N \rightarrow \infty .$$

The proof given in [3] will be sketched: it was based on a lemma similar to Lionville's theorem:

Lemma: If $\xi_1, \dots, \xi_k$ are linearly independent
irrationals in a real algebraic field of degree $n(\geq k+1)$,
there is a constant $C > 0$ such that if $q_1, \dots, q_k, p$
are any (rational) integers, not all zero,

$$(20) \quad |q_1 \xi_1 + \dots + q_k \xi_k - p| \geq \frac{C}{\max_{(j)} |q_j|^{n-1}} .$$

Proof. (Compare Cassels^[6], p.79). It is sufficient to prove the lemma in the case in which the $\xi_1, \dots, \xi_k$ are algebraic integers: for, in the general case, (20) can be multiplied through by a common denominator of the ξ 's (then, only integers appear on the left); if the resulting inequality holds, then, (20) also holds as written. Therefore, suppose the ξ 's are algebraic integers.

Denote the conjugates of ξ_j by $\xi_j^{(1)}, \dots, \xi_j^{(n)}$, where $\xi_j = \xi_j^{(1)}$, $j = 1, 2, \dots, k$. Construct the algebraic integer

$$\pi = q_1 \xi_1 + \dots + q_k \xi_k - p$$

and note that π and its conjugates are all $\neq 0$. Denote its conjugates by $\pi^{(1)}, \dots, \pi^{(n)}$, where $\pi = \pi^{(1)}$. The product of the conjugates,

$$N[\pi] = \pi^{(1)} \cdot \pi^{(2)} \cdot \dots \cdot \pi^{(n)},$$

is the norm of π and is, therefore, a rational integer.

Now, $N[\pi] \neq 0$; consequently

$$|N[\pi]| \geq 1$$

or

$$(21) \quad |\pi^{(1)}| \geq \frac{1}{|\pi^{(2)}| \cdot |\pi^{(3)}| \cdot \dots \cdot |\pi^{(n)}|}.$$

Now restrict the discussion to those sets of values of the rational integers $q_1, \dots, q_k, p$ for which

$$(22) \quad |\pi^{(1)}| = |q_1 \xi_1 + \dots + q_k \xi_k - p| < \frac{1}{2};$$

then,

$$(23) \quad |\pi^{(m)}| = |\pi^{(1)} + \sum_{j=1}^k q_j (\xi_j^{(m)} - \xi_j^{(1)})|$$

and, by the triangle inequality,

$$|\pi^{(m)}| \leq \frac{1}{2} + \sum_{j=1}^k |q_j| \cdot |\xi_j^{(m)} - \xi_j^{(1)}|.$$

Call

$$(24) \quad A = \frac{1}{2} + \max_{j,m} |\xi_j^{(m)} - \xi_j^{(1)}|.$$

Note that not only must one of the $q_1, \dots, q_k, p$ be different from zero, but one of the $q_1, \dots, q_k$ must be different from zero, because, otherwise, the inequality (22) could not hold.

Therefore, $\sum_j |q_j| \geq 1$; consequently, from (23),

$$(25) \quad |\pi^{(m)}| \leq \sum_{j=1}^k |q_j| \left(\frac{1}{2} + |\xi_j^{(m)} - \xi_j^{(1)}| \right) \leq A \sum_{j=1}^k |q_j|.$$

Note that, by definition (24), A is independent of the $q_1, \dots, q_k, p$. From (21) and (25),

$$|\pi^{(1)}| \geq \frac{1}{A^{n-1} \sum_{j=1}^k |q_j|^{n-1}}.$$

So far we have proved that either this inequality holds or $|\pi^{(1)}| \geq \frac{1}{2}$. Therefore, calling

$$C = \min \left\{ \frac{1}{(kA)^{n-1}}, \frac{1}{2} \right\},$$

we have the desired result:

$$|\pi^{(1)}| \geq \frac{C}{\max_{(j)} |q_j|^{n-1}}.$$

The proof of the theorem immediately preceding the lemma is then as follows: For any x , $\exp 2\pi i \{x\} = \exp 2\pi i x$; consequently,

$$(26) \quad I_N = \sum_{q_1, \dots, q_k} a(q_1, \dots, q_k) \frac{1}{N} \sum_{n=1}^N e^{2\pi i n(q_1 \xi_1 + \dots + q_k \xi_k)}.$$

From the formula for the Fourier coefficients, clearly $a(0, 0, \dots, 0) = I$. The factor multiplying $a(0, \dots, 0)$ in (26) is unity; therefore, we can form $I_N - I$ from (26) by merely omitting the term with $q_1 = q_2 = \dots = q_k = 0$. This omission will be denoted by a prime on the summation sign. That is

$$I_N - I = \frac{1}{N} \sum'_{q_1, \dots, q_k} a(q_1, \dots, q_k) \frac{e^{2\pi i N(q_1 \xi_1 + \dots + q_k \xi_k)} - 1}{1 - e^{-2\pi i (q_1 \xi_1 + \dots + q_k \xi_k)}},$$

or, by the triangle inequality,

$$(27) \quad |I_N - I| \leq \frac{1}{N} \sum'_{q_1, \dots, q_k} |a(q_1, \dots, q_k)| \left| \frac{\sin \pi N(q_1 \xi_1 + \dots + q_k \xi_k)}{\sin \pi (q_1 \xi_1 + \dots + q_k \xi_k)} \right|.$$

For any real x ,

$$|\sin \pi N x| \leq 1,$$

$$|\sin \pi x| \geq 2|x - \langle x \rangle|,$$

where $\langle x \rangle$ denotes the integer nearest to x ; and therefore, finally

$$|I_N - I| \leq \frac{1}{2N} \sum_{q_1, \dots, q_k}' \frac{|a(q_1, \dots, q_k)|}{|q_1 \xi_1 + \dots + q_k \xi_k - p|},$$

where p denotes the integer nearest to $q_1 \xi_1 + \dots + q_k \xi_k$; or by the lemma,

$$|I_N - I| \leq \frac{1}{2CN} \sum_{q_1, \dots, q_k}' \left\{ \max_{(j)} |q_j| \right\}^{n-1} |a(q_1, \dots, q_k)|;$$

hence, by (19), $I_N - I = O(\frac{1}{N})$, as stated.

The practical applicability of the theorem is limited by the fact that the inequality (19) imposes drastic restrictions on the integrand $f(x_1, \dots, x_k)$: f and all its partial derivatives up to order $n - 1$ must be smooth (possess absolutely convergent Fourier series). Still, the theorem indicates that the sequence of points generated in the unit cube by (6) is more uniform, in some sense, than a random sequence would be. One seems therefore justified in expecting that, even with not-so-smooth integrands, the error $I_N - I$ may decrease more rapidly than $O(1/\sqrt{N})$, although perhaps not as rapidly as $O(1/N)$.

The result was greatly improved by L. G. Peck, who proved the following theorem^[4]: If $\xi_1, \dots, \xi_k$ are in a real algebraic field of degree $k + 1$, and if the Fourier coefficients of $f(x_1, \dots, x_k)$ satisfy the inequality

$$(28) \quad |a(q_1, \dots, q_k)| < \frac{C}{\max_{(j)} |q_j|^{k+\epsilon}}$$

for some $C < \infty$ and some $\epsilon > 0$, then $I_N - I = O(\frac{1}{N})$.

To compare the two theorems, we may take the case $n = k + 1$ in (19). In order that an inequality of the form (19) imply (28), there is required only an infinitesimal increase in the exponent of (19); whereas, in order that an inequality of the form (28) imply (19), there is required an increase by $k - 1$ of the exponent of (28). Therefore, (28) covers essentially a wider class of functions than (19); unfortunately, functions having a discontinuity of value or of low order partial derivatives are still not covered, when k is large.

4. Statistical Treatment.

With I and I_N defined by (1), (2), (6), we now regard $\xi_1, \dots, \xi_k$ as random variables, uniformly and independently distributed between 0 and 1, and we consider the distribution of the values of I_N , for fixed N and a fixed function $f(x_1, \dots, x_k)$. We shall prove: if the Fourier series

$$f(x_1, \dots, x_k) = \sum_{-\infty}^{\infty} (q_1, \dots, q_k) a(q_1, \dots, q_k) e^{2\pi i(q_1 x_1 + \dots + q_k x_k)}$$

is absolutely convergent; that is, if

$$(29) \quad \sum_{-\infty}^{\infty} (q_1, \dots, q_k) |a(q_1, \dots, q_k)| = B < \infty ,$$

then,

$$(30) \quad \langle I_N \rangle = I ,$$

$$(31) \quad \langle |I_N - I| \rangle \leq \frac{1 + \ln N}{N} B ,$$

where $\langle \rangle$ denotes expectation value.

To prove (30), note that the expectation value of each term of (26) vanishes, except for the terms for which $q_1 = \dots = q_k = 0$; and these terms give (30) because $a(0, \dots, 0) = I$. The term-wise integration is justified by the absolute convergence.

To prove (31), we start from (27), each side of which is to be integrated over the k -dimensional unit cube $0 \leq \xi_1 < 1, \dots, 0 \leq \xi_k < 1$. Because of the absolute convergence, and because of the positivity, we can equally well integrate any given term of the series over any other region (say, rectangular), of unit volume, that includes a complete period of the variation. For given values of $q_1, \dots, q_k$, consider the slab

$$-\frac{1}{2} \leq q_1 \xi_1 + \dots + q_k \xi_k < \frac{1}{2}$$

in the k -dimensional space $\xi_1, \dots, \xi_k$. If we take an appropriately oriented cartesian coordinate system, the corresponding term in (27) depends only on the coordinate

normal to the faces of this slab; furthermore, the thickness of the slab is equal to the period. We may therefore integrate over a rectangular region consisting of a piece of this slab having cross-section area equal to the reciprocal of its thickness. Denoting this region by P , we have

$$\begin{aligned} & \int \cdots \int_P \left| \frac{\sin \pi N (q_1 \xi_1 + \cdots + q_k \xi_k)}{\sin \pi (q_1 \xi_1 + \cdots + q_k \xi_k)} \right| d\xi_1 \cdots d\xi_k \\ &= \int_{-1/2}^{1/2} \left| \frac{\sin \pi N y}{\sin \pi y} \right| dy . \end{aligned}$$

To estimate this, we split the integral into 2 parts: for $|y| < \frac{1}{2N}$, we use the bound $\left| \frac{\sin \pi N y}{\sin \pi y} \right| \leq N$; for $\frac{1}{2N} \leq |y| \leq 1$, we use $|\sin \pi N y| \leq 1$ and $|\sin \pi y| \geq 2|y|$; the above integral is then less than

$$\begin{aligned} & \frac{1}{2N} \int_{-1/2}^{1/2} N dy + 2 \int_{\frac{1}{2N}}^{\frac{1}{2}} \frac{dy}{2y} = 1 + \ln N . \end{aligned}$$

therefore, $\langle |I_N - I| \rangle \leq \frac{1 + \ln N}{N} B$, as stated.

To make use of this result, one would presumably make a number, say, M , of random choices of the vector $(\xi_1, \dots, \xi_k)$; for each of these choices, one would generate N chains of events, as described in Section 1, using Weyl's sequence (6) (there are then NM chains in all). One would then have M values of I_N ;

their average would be taken as an estimate of I and their dispersion as a measure of the probable error. The numerical examples in Section 6 show that the resulting mean error is substantially less than would be expected if the NM chains had been chosen at random.

However, the foregoing theorem leaves unanswered various questions about the distribution of the values of I_N , for fixed N , as the $\xi_1, \dots, \xi_k$ vary randomly; there is no indication in the above that this distribution is even approximately normal, even for large N . Therefore, in particular, the dependence of the standard deviation $\sigma = \sqrt{\langle (I_N - I)^2 \rangle}$ on N is an open question. [If the distribution were normal, the standard deviation would be equal to $\sqrt{\frac{\pi}{2}}$ times the first absolute moment $\langle |I_N - I| \rangle$ and hence it would also vary as $O(\frac{1 + \ln N}{N})$.] This question is of some slight practical importance, because the probable error of the sample mean (mean of the values of I_N) depends, according to the central-limit theorem, on σ rather than on $\langle |I_N - I| \rangle$ (σ exists and, in fact, the distribution is bounded, so that all its moments exist); but an error estimate based* on $\langle |I_N - I| \rangle$ is probably good enough for practical purposes.

* i.e., using the formula that would be correct if the distribution of I_N were normal.

5. Comments.

Some miscellaneous comments are collected in this section.

a. Number of digits needed. In a computing machine the $\xi_1, \dots, \xi_k$ are approximated by rational numbers (binary or decimal fractions); the question arises as to the number of digits that must be retained in order that the error due to this source is no greater than the inherent error of the method, for given N . The answer is, roughly, that slightly more than $2 \log_{10} N$ decimal places or $2 \log_2 N$ binary places should suffice. To see this, suppose that each of the $\xi_1, \dots, \xi_k$ is in error by an amount of order $\delta\xi$. Then, a given component of the vector $(x_1^{(n)}, \dots, x_k^{(n)})$, say $x_j^{(n)}$, that is, $\{n\xi_j\}$, is in error by an amount of order $n\delta\xi \leq N\delta\xi$. The resulting error of $f(x_1^{(n)}, \dots, x_k^{(n)})$ is of order $\leq KN\delta\xi$, where*

$$K = \text{Max}_{(x_1, \dots, x_k)} \text{Max}_{(j)} \left| \frac{\partial f(x_1, \dots, x_k)}{\partial x_j} \right| .$$

Each of the k components of $(x_1, \dots, x_k)$ can contribute such an error, and these errors must be compounded. However, $f(x_1, \dots, x_k)$ does not generally depend on all k of its arguments, because k is chosen, as stated in Section 1, to be much larger than the expected number of "random numbers"

* In the examples, $f(x_1, \dots, x_k)$ is continuous and has piecewise continuous, hence k bounded, first partial derivatives.

needed to construct an individual chain. We may define an effective number of dimensions, depending on the function $f(x_1, \dots, x_k)$, as follows: as the coordinates $x_1, \dots$ of a given point in the cube are examined, in succession, in the construction of a chain, it is found that only a certain number of them, say $x_1, \dots, x_{k^*}$ are needed to construct that chain. In the neighborhood of this point in the cube, $f(x_1, \dots, x_k)$ is thus actually independent of the remaining variables $x_{k^*+1}, \dots, x_k$ (generally $k^* \ll k$). The effective number of dimensions k_0 is then defined as the average of k^* over the unit cube. k_0 is of order 10 for simple Monte-Carlo problems, even though many more of the variables $x_1, \dots, x_k$ may occasionally be needed for the construction of an exceptional chain. Therefore, the error of a single value of $f(x_1, \dots, x_k)$ is of order $\leq k_0 \text{ KN} \delta \xi$. The error of the average (2) is also $\leq k_0 \text{ KN} \delta \xi$, even if no allowance is made for possible mutual cancellation. On the other hand, the inherent error of the method can hardly, in any case, be less than about K/N (see next paragraph), hence we want

$$k_0 \text{ N} \delta \xi \leq \frac{1}{N} ,$$

or, if the $\xi_1, \dots, \xi_k$ are themselves of order unity,

$$\log_{10} \frac{1}{\delta \xi} \geq 2 \log_{10} N + \log_{10} k_0 ,$$

(32)

$$\log_2 \frac{1}{\delta \xi} \geq 2 \log_2 N + \log_2 k_0 .$$

In the examples, with $N \sim 10^4$, $k_0 \sim 10$, it follows that nine decimal places should suffice; eleven were used.

b. Comparative rates of convergence of various procedures. In any method in which there is a sequence of points in the cube and the integral is approximated by the average (equation (2)) the error is certain to be $\geq \mathcal{O}(\frac{1}{N})$, unless there is some rule or procedure for stopping the sequence at those exceptionally fortunate values of N where the error is smaller, because the difference between successive approximations I_N and I_{N+1} can be of order $\frac{1}{2} R \frac{1}{N}$, where $R = \text{range}^*$ of $f(x_1, \dots, x_k) = \text{maximum of } f(x_1, \dots, x_k) - \text{minimum of } f(x_1, \dots, x_k)$. In any method of this kind that gives this small an error, one may say, roughly, that errors introduced early in the sequence are eventually almost completely compensated, leaving a residual that is not appreciably larger than the error of the last term. Hence, the estimates (16), (18), (31) are about as good as could be hoped for.

* In the preceding section, this lower bound of the error was used; there, however, we wrote $K \frac{1}{2}$ (maximum magnitude of the 1st derivatives) instead of $\frac{1}{2} R$; in practical cases, these will be of the same order of magnitude.

It is interesting to compare the present method and the Monte-Carlo method, regarded as methods for evaluating multi-dimensional integrals, with methods using a lattice of points - say, the trapezoid-rule method*. If each of the k edges of the unit cube is divided into m intervals, there will be m^k points in the lattice**. The trapezoid rule gives then just the formula (2), but with N restricted to values of the form m^k . The error of the trapezoid rule formula is $\mathcal{O}(\Delta x) = \mathcal{O}(\frac{1}{m})$; i.e.,

$$I_N - I = \mathcal{O}\left(\frac{1}{N^{1/k}}\right),$$

compared with $\mathcal{O}(1/\sqrt{N})$ for random sampling. In three or more dimensions, the Monte-Carlo method is better than the trapezoid rule, at least for large N . By means of more elaborate quadrature formulas, using a lattice one can achieve $\mathcal{O}((\Delta x)^2)$ for numerical integration of smooth functions, i.e.

$$I_N - I = \mathcal{O}\left(\frac{1}{N^{2/k}}\right).$$

* For periodic functions, the trapezoid rule is preferable to Simpson's rule.

**Corresponding points on opposite faces of the cube may be identified, because $f(x_1, \dots, x_k)$ is periodic; then, all m^k points have equal weight in the numerical integration.

Still, Monte Carlo is better for large N , in five or more dimensions. The method of systematic sampling is asymptotically better than the trapezoid rule already in two dimensions and better than second-order quadrature formulas in three dimensions.

c. Conjectured result for algebraic numbers. It is conjectured that, for reasonable^{*} integrands, if $\xi_1, \dots, \xi_k$ are algebraic, and if $\epsilon > 0$, then,

$$(33) \quad I_N - I = \mathcal{O}\left(\frac{1}{N^{1-\epsilon}}\right) .$$

It was observed, in Section 3, that this is true in one dimension ($k=1$), by combining a result of Ostrowski with Roth's theorem. But more can be said: it has been proved that this is always true for sufficiently smooth integrands (for example analytic) and one can show that the introduction of certain discontinuities does not disturb it. We can, in fact, add to the smooth function any number of functions, each depending only on some linear combination of the coordinates, that is, of the form

$$f(x_1, \dots, x_k) = g(a_1 x_1 + \dots + a_k x_k)$$

* If the reader insists on a more definite statement of the conjecture, let us say either that $f(x_1, \dots, x_k)$ is piecewise differentiable or continuous and piecewise differentiable.

where the $a_1, \dots, a_k$ are rational^{*} coefficients, and where $g(w)$ is smooth except, say, for a finite number of simple discontinuities. This is easily seen, because, if $f(x_1, \dots, x_k)$ is periodic, we can just as well write $x_j^{(n)} = n\xi_j$ instead of $x_j^{(n)} = \{n\xi_j\}$, so that

$$a_1 x_1 + \dots + a_k x_k = n(a_1 \xi_1 + \dots + a_k \xi_k) ;$$

and the integral of g therefore reduces to the one-dimensional case, $a_1 \xi_1 + \dots + a_k \xi_k$ being an algebraic number.

We have not been able to prove or disprove the above conjecture.

d. Special choices of the algebraic irrationals.

If, as suggested by the above conjecture, a favorable limitation of the error can be obtained using a single set of algebraic numbers for $\xi_1, \dots, \xi_k$, the question arises how to choose these numbers to give the best results.

J. von Neumann suggested to the writer, many years ago, that one should try to obtain results based on a particular choice, rather than general theorems of the kind given in Section 3. For instance, the reasoning in Section 3 suggests that the $\xi_1, \dots, \xi_k$ should be so chosen that the diophantine inequalities

$$|\xi - \frac{p}{q}| \geq \dots$$

* they could also be algebraic.

are as strong as possible. Since, for the best approximants, according to the theory of continued fractions

$$\left| \xi - \frac{p_1}{q_1} \right| \leq \frac{1}{q_1 q_{1+1}} ,$$

it seems that the ξ 's should be so chosen that the q_i increase as slowly as possible. Therefore, the partial denominators of the continued-fraction expansions should be as small as possible. Some numerical experimentation has been performed, with $\xi_1, \dots, \xi_k$ so chosen that their continued-fraction expansions have only 1's and 2's as partial denominators. The results did not seem to be noticeably better than with other choices that have been tried. The reason may be that, even if ξ_1 and ξ_2 are two such numbers, there is no apparent reason for supposing that $\xi_1 + \xi_2$, or other simple linear combinations, will also have small partial denominators.

3. An alternative to I_N as estimator of I . In equation (2), which expresses I_N as an average of the values of the function at N points in the cube, all N points are given equal weight. In random sampling, this is appropriate, because the points are statistically independent, hence uncorrelated. In systematic sampling, there are correlations, in fact, long-range ones, among the points of the sequence; conceivable, unequal weighting

might be better. The following heuristic arguments lead to a set of unequal weights.

If the successive values of

$$S_m = \sum_{l=1}^m (n) f(x_1^{(n)}, \dots, x_k^{(n)})$$

are plotted, on a graph, against m , the points on the graph approximate a straight line whose asymptotic slope is equal to I . The problem is to estimate the slope from a finite number of points.

In random sampling, the successive increments $S_{m+1} - S_m$ in the graph have equal probable errors, and are uncorrelated; in this case, the best estimate of the slope is obtained by using only the first and the last points of the graph: this gives equal weight to all the increments.

The estimates of Sections 3 and 4 give $I_N - I$ much closer to $O(\frac{1}{N})$ than to $O(\frac{1}{\sqrt{N}})$; and this suggests that, for systematic sampling, it is the actual coordinates of the points in the graph, rather than the increments, that have equal probable error. If this is so, one should get a better estimate of the slope of the graph by a least-squares fitting. This leads to the formula

$$(34) \quad \tilde{I}_N = \frac{1}{N} \sum_{l=1}^N (n) \frac{6n(N+1-n)}{(N+1)(N+2)} f(x_1^{(n)}, \dots, x_k^{(n)})$$

in place of I_N . One can still prove that

$$\langle \tilde{I}_N \rangle = I$$

$$\langle |\tilde{I}_N - I| \rangle = \left(\frac{\ln N}{N} \right)$$

(although the proof is harder) when the $\xi_1, \dots, \xi_k$ are treated statistically, as in Section 4, if $f(x_1, \dots, x_k)$ has an absolutely convergent Fourier series. In the numerical examples, the improvement in going from I_N to $\tilde{I}_N$ was disappointingly small.

6. Numerical Examples.

By way of test, the method was applied to the following simple multiplicative problem. At time $t = 0$, a single particle is present; at any instant, any particle can divide into two particles; the probability of division of a given particle in an interval Δt is taken as constant $\lambda \Delta t$; and, in fact, the constant is set equal to 1 by suitable choice of the unit of time; no other processes are considered; the only quantity of interest is the number $m(t)$ of particles present, as function of the time t . This problem can be treated by elementary methods, according to which

$$\langle m(t) \rangle = e^t$$

$$\langle (m(t) - e^t)^2 \rangle = e^{2t} - e^t ;$$

and, in fact, if $p_m(t)$ is the probability of having just m particles present, at time t , then $p_m(t)$ is the coefficient of x^m in the expansion of the generating function

$$\phi(x,t) = \frac{x}{x+(1-x)e^{-t}} \quad .$$

We shall construct each chain, up to a time $t = T$, called the census time, and we wish to estimate the expected number $\langle m(T) \rangle$ of particles present at time $t = T$.

When the random numbers $x_1, x_2, \dots, x_k$, i.e. the coordinates of a point in the k -dimensional cube, are available, the construction of a chain proceeds as follows. We adopt the convention (purely for purposes of terminology) of saying that, when a particle divides, it dies or disappears and two new particles are born. We keep a running list (stored in the machine) of the times of birth of all particles whose times of death we have not yet ascertained. At the beginning of the calculation of this chain, the list contains just one entry, namely the time $t = 0$ for the particle present at the beginning. We examine the entries of the list, one by one, and ascertain the time of death by adding* $-\ln x$ to

* In the actual machine work, we store e^{-t} throughout, instead of t , to avoid the calculation of logarithms; we then multiply e^{-t} by x .

the time of birth, where x is the next unused random number of the sequence $x_1, \dots, x_k$. Thus the probability of the particle's survival to time t is $\exp -(t - \text{time of birth})$. If this time of death is $\geq T$ we simply strike the particle off the list; if it is $< T$ we also strike it off, but we then add two new particles to the list, with birth times equal to the time of death of the old particle. This procedure is continued until the list is exhausted; the description of this particular chain is then complete. The crude, simple, way to estimate $\langle m(T) \rangle$ would be to take $f(x_1, \dots, x_k)$ equal to the actual number of particles $m(T)$ present in the chain at time $t = T$. This is equal to the total number of particles struck off the list during the procedure just described, and can be counted by the machine during the construction of the chain. In this case, $f(x_1, \dots, x_k)$ is a step function with integer values.

To find a smoother function of the $x_1, \dots, x_k$ as estimator of $\langle m(T) \rangle$, we proceed as follows. If $d(t)$ is the number of divisions that have occurred, up to time t , then, clearly,

$$\begin{aligned} m(t) &= 1 + d(t) , \\ \langle m(T) \rangle &= 1 + \langle d(T) \rangle , \end{aligned}$$

because each division increases the number of surviving particles by 1. We estimate $d(T)$ by looking at the birth

times (including $t = 0$ for the original particle) and computing, for each particle, the probability that a particle born at that time will die (i.e. divide) before $t = T$. The sum of these probabilities is

$$(35) \quad \tilde{d}(T) = \sum_{(i)} [1 - e^{-(T-t_i)}] ,$$

where the sum is over all the particles in the chain, and t_i is the time of birth of the i^{th} particle. Clearly, $\tilde{d}(T)$ is a continuous, piecewise differentiable function of $x_1, \dots, x_k$; and

$$\langle \tilde{d}(T) \rangle = \langle d(T) \rangle .$$

$f(x_1, \dots, x_k)$ is taken as $1 + \tilde{d}(T)$. This choice of $f(x_1, \dots, x_k)$ is an example of a technique, called the use of expected values by Kahn^[7], for reducing the variance in Monte Carlo calculations.

The actual construction of the chains is the same as described above except for one modification: in order that $f(x_1, \dots, x_k)$ have an absolutely convergent Fourier series, it does not suffice that it merely be continuous and piecewise differentiable in the unit cube; it should also have equal values at corresponding points on opposite faces of the cube, so that the periodic function given by the Fourier series is

continuous in the entire k -dimensional space. To achieve this is, however, very easy: whenever a random number is needed, one uses $|2x-1|$ instead of x itself, where x is the next unused member of the sequence $x_1, \dots, x_k$. $|2x-1|$ is also uniformly distributed in $(0,1)$, if x is; and, by this trick, $f(x_1, \dots, x_k)$ is symmetrized about the mid-planes of the cube.

In order to compare our results with those of random sampling (Monte Carlo), we must have an estimate of the probable error of the random sampling, using the same function $f(x_1, \dots, x_k)$. (If we use the expected-values technique described above, it is only fair to permit the same in the Monte Carlo method, whereby its error is somewhat reduced.) This is given by

$$\text{p.e.} = 0.6745 \frac{\sigma}{\sqrt{N}},$$

where

$$(36) \quad \sigma = \sqrt{\langle f^2 \rangle - \langle f \rangle^2}.$$

The expectation value $\langle f^2 \rangle$ is rather difficult to compute analytically, in the present problem, but an adequate (in fact, better-than-adequate) estimate can be obtained, concurrently with that of $\langle f \rangle$, by the computing machine, using the same method (i.e., the systematic sampling). This we have

done, in the examples.

In a first series of tests, the $\xi_1, \dots, \xi_k$ were taken as the fractional parts of the numbers

$$\sqrt{2}, \sqrt{3}, \sqrt{6}, \sqrt{5}, \sqrt{10}, \sqrt{15}, \sqrt{30}, \sqrt{7}, \dots$$

The radicands are square-free integers, taken in an order that is easy for the machine to generate. It is known that the above are linearly independent irrationals.

Figure 1 shows values of the error $\tilde{I}_N - I$ for several problems. In each problem, there were $N = 21,000$ chains*. The problems have different values of the census time T ; the abscissa is the quantity e^T , which, as noted above, is the expected number of particles at $t = T$ and is equal to I . The curves give the probable error of the corresponding calculation by random sampling; on a statistical basis, half the points should lie between the two curves, and half outside.

It is seen that the error is reduced, relative to random sampling, by a factor that ranges from about $1/2$ to about $1/35$ (the last case is probably fortuitously good). By comparison, the statistical treatment of Section 4 (there is, of course, no reason why it should apply here, since a single choice of the numbers $\xi_1, \dots, \xi_k$ has been made here) would suggest an

* This number was determined by certain limitations of the computation program that are probably of no interest to the reader.

improvement by a factor of the order $((1+\ln N)/N)/\sqrt{N} \approx 1/13$.

Figure 2 shows the results of some calculations by the statistical method of Section 4, all for the case $e^T = 2$. Five values of N were taken: $N = 1000, 2000, 4000, 10,000, 21,000$. Each point represents the results of 37 different calculations ($M = 37$), with $\xi_1, \dots, \xi_k$ obtained from a random number generator that has been used and tested in Monte Carlo work. What is plotted is the average of the 37 values of $|\tilde{I}_N - I|$ thus obtained ($I = 2$). Although $M = 37$ represents a rather small sample, this average should not differ too much from the expectation value $\langle |\tilde{I}_N - I| \rangle$ discussed in the theorem. The data are too meager to tell whether a $(1+\ln N)/N$ law is followed; and, besides, the theorem gives only an inequality; furthermore, the value of constant in the theorem is not known. However, if random sampling had been used, and if the sample size (i.e. number of calculations, $M = 37$) is not too small, the points would lie on the curve shown; hence, an improvement over random sampling is again apparent.

Figure 3 shows similar results for the case $e^T = 4$. The improvement over random sampling is somewhat less than for $e^T = 2$.

It is not known whether an improvement of this sort can be expected in practical Monte Carlo problems. It should perhaps be pointed out that the incorporation of systematic

sampling in a Monte Carlo code is not at all difficult; the increase (if any) in computing time is negligible; and the increase in required storage is not severe - a few hundred memory locations are needed for the $\xi_1, \dots, \xi_k$ and, perhaps, for the current values of $x_1^{(n)}, \dots, x_k^{(n)}$.

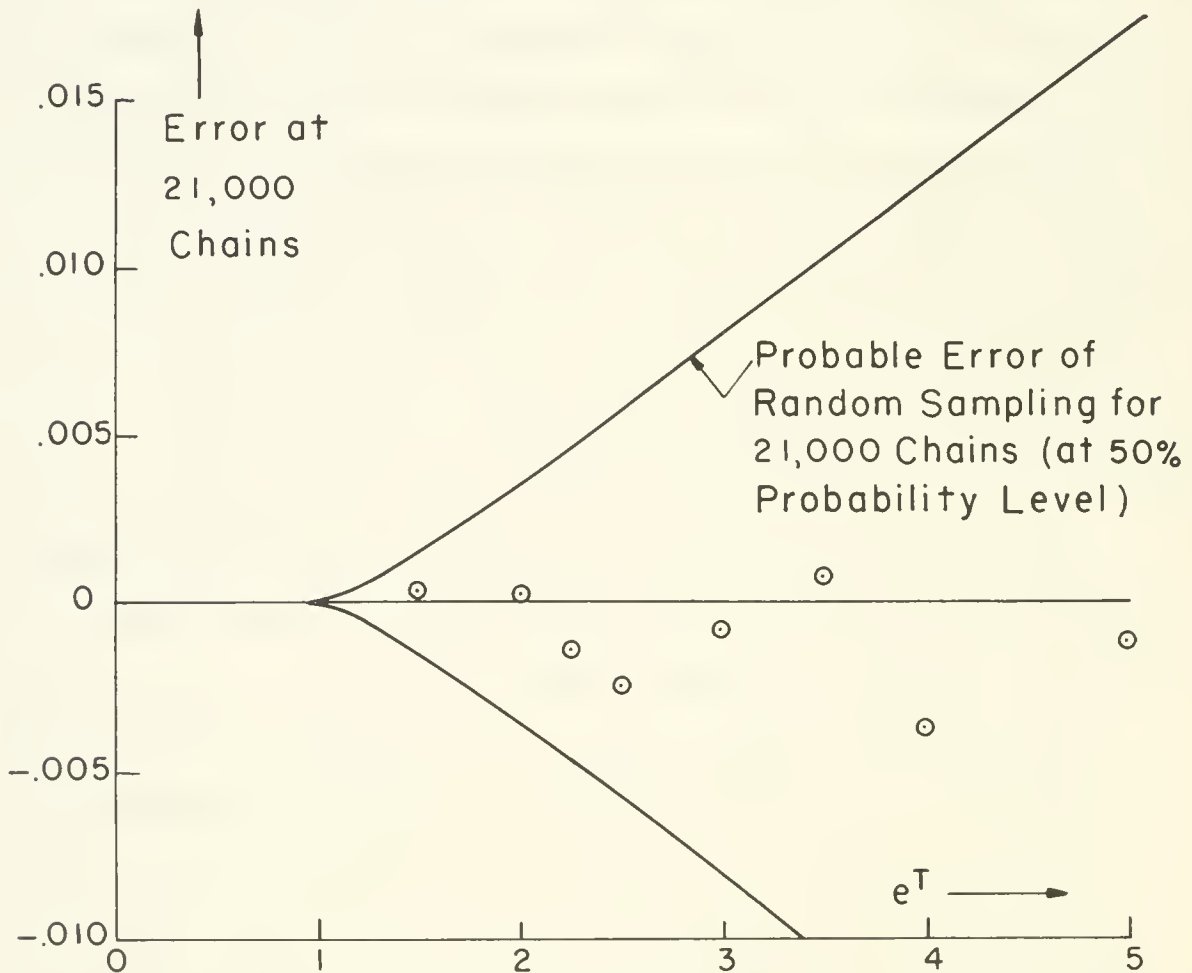


FIG . 1

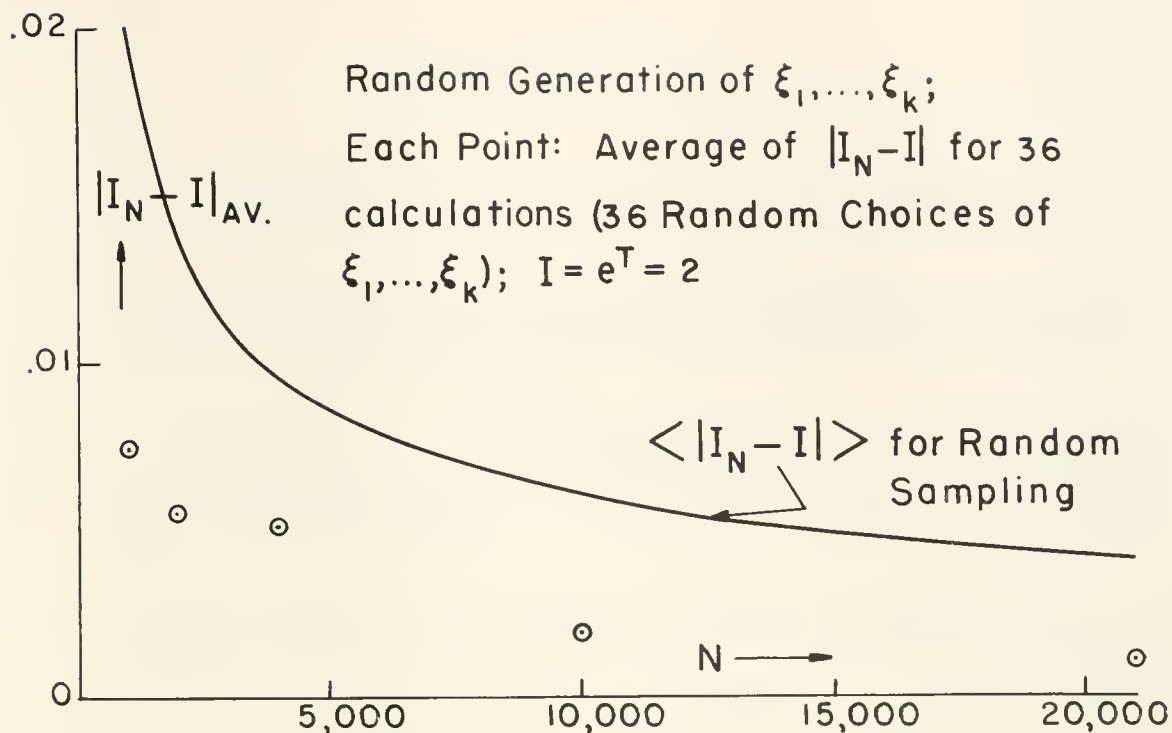


FIG. 2

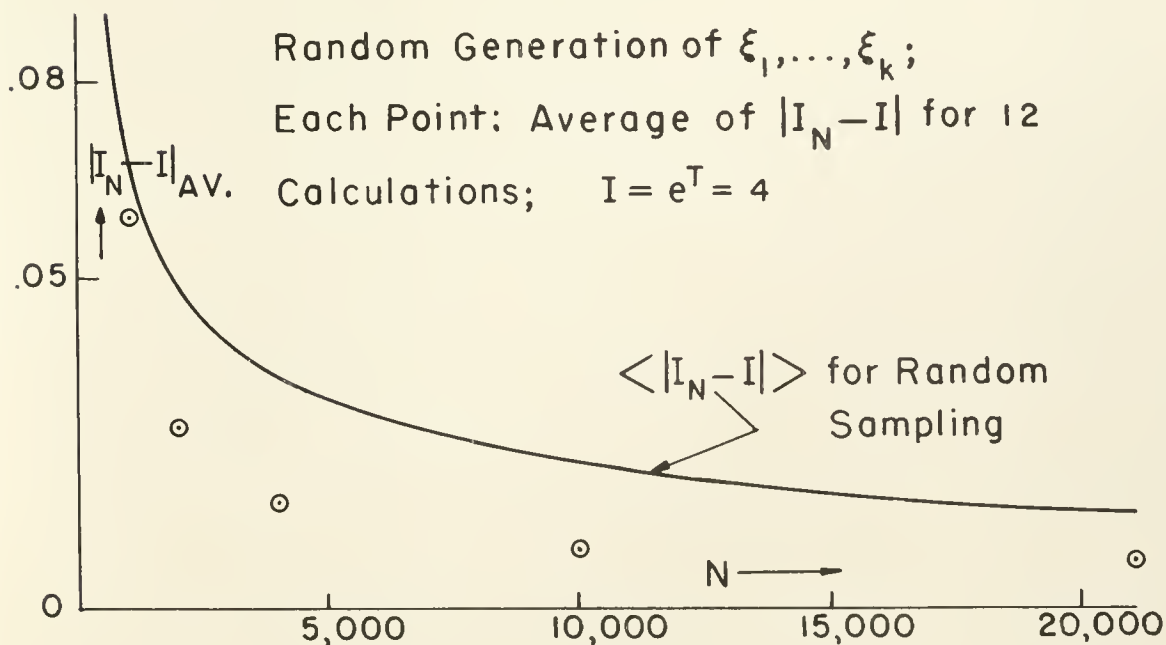


FIG. 3

BIBLIOGRAPHY

- [1] Weyl, H., Über die Gleichverteilung von Zahlen mod. Eins., Math. Annalen, vol. 77, pp. 313-352 (1916)
- [2] Ostrowski, A., Bemerkungen zur Theorie der Diophantischen Approximationen, Abhandlungen aus dem Mathematischen Seminar der Hamburgischen Universität, vol. 1, pp. 77-98 (1922)
- [3] Richtmyer, R. D., The Evaluation of Definite Integrals, and a Quasi-Monte-Carlo Method based on the Properties of Algebraic Numbers. Los Alamos Scientific Laboratory Report LA-1342 (1951)
- [4] Peck, L. G., On Uniform Distribution of Algebraic Numbers. Proc. Amer. Math. Soc. vol. 4, pp. 440-443 (1953)
- [5] Roth, K. F., Rational Approximations to Algebraic Numbers. Mathematika, vol. 2, pp. 1-20 and 168 (1955)
- [6] Cassels, J. W. S., Introduction to Diophantine Approximations. Cambr. Univ. Press (1957)
- [7] Kahn, H., Applications of Monte Carlo. The RAND Corporation, Santa Monica, Calif., Report RM-1237-AEC, April 1954

NYO
8674 Richtmyer
A non-random
sampling method...

NYU
NYO
8674 Richtmyer
AUTHOR
A non-random sampling method
based on congruences...

DATE DUE	BORROWER'S NAME	ROOM NUMBER
JUL 17 1975	M. Shimazaki	
162 Street oil due 6/11/77		
MAY 31 1985	SHERMAN	

N. Y. U. Institute of
Mathematical Sciences
25 Waverly Place
New York 3, N. Y.

JAN 22 1985

Date Due

JUL 17 1975

162 Street oil due 6/11/77

MAY 31 1985

MAY 12 1987

MAY 31 1987

MAY 27 1987

ILL - 5-28-88

